Environmental Health Perspectives
Vol. 61, pp. 239-255, 1985

Substructural QSAR Approaches and Topological Pharmacophores

by Rainer Franke*, Stefan Huebel* and W. Juergen Streich*

For large and diverse data sets, simple QSAR methods based on linear and additive models can no longer be applied. In such cases topological methods using descriptors directly derivable from two-dimensional chemical structures provide a useful alternative. The results of such analyses can be used for lead optimization, to guide biological testing and even aid in the design of novel compounds. Various types of topological descriptors and algorithms are briefly discussed. Which of those is to be selected depends on the objective of the investigation and the properties of the data set. Two new methods, LOGANA and LOCON, are discussed in some more detail. With the help of these methods, substructural patterns ("topological pharmacophores") characteristic of compounds possessing a certain biological property can be evaluated. Both methods are designed in such a way that full use can be made of the data handling capacity of computers while maintaining an optimal impact of the experience of the researcher. They are model-free and do not require any mathematical knowledge. While LOGANA deals with semiquantitative or even qualitative biological data, LOCON can be applied to activity data on a continuous scale. The basic procedure in both cases consists in the stepwise combination of substructural descriptors by the logical operations "and," "or" and "not." With a simple example the utility of the methods is demonstrated.

Introduction

Quantitative structure–activity relationships (QSARs) have become an indispensable tool to rationalize the interaction of chemical compounds with living matter. The basic philosophy of QSAR methods is to draw conclusions by analogy assuming that similarity of drugs with respect to certain chemical properties will result in similar biological responses. The problem, then, is to determine what these properties are and how they are connected with the biological activity of interest. To this end a set of compounds with known biological activities (which will be called a "training series" throughout this paper) is analyzed. The principal steps are always roughly the same. First, a set of chemical descriptors is selected so that all chemical properties of the compounds that may be important for their biological action are believed to be adequately characterized. The values of all these descriptors are then collected or evaluated for all compounds of the training series and fed into a computer together with the corresponding biological activity data. The computer compares and connects the descriptors and activities by means of a suitable algorithm until a QSAR is found. In the broadest sense, a QSAR may be regarded as a computer-derived rule which quantitatively describes biological activity in terms of chemical descriptors. Once a QSAR is known, the following may become possible: conclusions and hypotheses as to molecular mechanisms of action; optimizations of a given lead compound that includes maximizing desired and minimizing undesired (e.g., toxic) biological effects; prediction of what kind of biological activity a new or as yet untested compound is likely to possess (which may aid in the preselection of compounds for screening programs in order to increase the incidence of actives, but also in recognizing potentially toxic chemicals); generation of new lead compounds.

Different approaches (and data sets) are required depending on which of the above objectives is in the foreground. A variety of QSAR methods has, in fact, been developed during the last years not only for this reason but also since the data to be handled may be very different. The training series, for example, can largely vary in size and with respect to the complexity or diversity of chemical structure variation. Biological activity data, on the other hand, may come from very different sources or tests and be expressed on many different scales. Thus, a whole tool box of methods for both the evaluation of descriptors as well as the application of computer algorithms is necessary to be able to cope with the many possible situations (*1–15*).

*Academy of Sciences of the GDR, Research Center of Molecular Biology and Medicine, Institute of Drug Research, Berlin, German Democratic Republic.

In this paper we wish to deal with the use of topological (substructural) descriptors in QSAR work. After a short review of currently used substructural approaches two new techniques for the evaluation of topological activity patterns, LOGANA and LOGON, will be discussed in more detail and illustrated with a simple example.

Why Topological Descriptors?

The most popular and widely used QSAR method these days certainly is the Hansch approach. Its principle consists in correlating the logarithm of biological activity (log A) with hydrophobic, electronic, and steric molecular parameters by means of linear multiple regression analysis. The result is equations in which log A as dependent variable is expressed by a weighted linear combination of those molecular parameters that turn out to make a statistically significant contribution to "explaining" the variance of log A in the training series. This approach requires precise enough biological activity data on a continuous scale, the applicability of linear free energy relationships that are based on a simple linear and additive model, and the availability of the respective molecular parameters for all structural variants in the training series. Usually, the latter conditions are only fulfilled in so-called congeneric series where at a constant parent structure only substituents are varied. But even there the model assumptions of Hansch analysis may break down if the substituent exchanges alter the structure too drastically (*16*), and for too exotic substituents appropriate parameters may not be available. One cannot but admire the ingenuity with which especially the Hansch group has pushed this type of analysis almost beyond its limits handling training series with several hundreds of compounds and partly very diverse structures in one analysis. In order to be able to do that, however, so-called "indicator or dummy" variables frequently become necessary in order to characterize structural variations that cannot adequately be accounted for by physicochemical parameters. These indicator variables are binary quantities usually describing the presence or absence of certain structural features according to

$$x_{ij} = \begin{cases} 1, & \text{if the } i\text{th feature is present in the } j\text{th compound} \\ 0, & \text{if not} \end{cases} \qquad (1)$$

In that sense, they already represent the simplest case of topological descriptors since they are directly derived from the topology (two-dimensional chemical formula) of the compounds under consideration.

QSAR equations containing both, terms with physicochemical and with indicator variables, represent a mixed form of the Hansch approach and the Free-Wilson analysis (*1,4,7,11,15*). The two methods are formally equivalent since, in both instances, the logarithm of biological activity is built up additively on the basis of a linear model from contributions of the substituents. The difference is that the Free-Wilson analysis does not require physicochemical molecule parameters but is entirely based on topological descriptors according to the model

$$\log A_j = \sum_i x_{ij} z_j + \mu \qquad (2)$$

where A_j is the biological activity of the j-th compound, x_{ij} is a topological descriptor according to Eq. (1) describing the absence or presence of the i-th substituent variable (which, in this case, includes type and position of substituent), z_j the contribution of the i-th substituent variant to log A_i, and μ is the contribution of a constant parent moiety. Free-Wilson and Hansch analyses supplement each other in that the former is particularly suitable in cases where relatively few substituents occur in many positions of substitution while the domain of the latter is where many substituents are varied in only few positions. The applicability of Free-Wilson analysis can be extended by introducing terms to account for intramolecular interactions (*15*) or by including fusion points or heteroatoms, etc., as variables (*17*). However, both Free-Wilson and Hansch analyses will break down or cannot be applied in either of the following cases: the biological measurements available are only semiquantitative or qualitative; or the structural changes in the starting set are so drastic that the linear model on which both methods are based is no longer valid. If structural variations become too extensive in really heterogeneous and large data sets one has to look for other possibilities. One way to go could be the application of molecular modeling techniques combined with interactive computer graphics. These methods, however, need precise biological information largely free from pharmacokinetic contributions and cannot be applied to large data sets. In addition, a certain amount of *a priori* knowledge and a hypothesis to start with are usually required. This brings us back to topological descriptors which are easy to derive for any structure but must, of course be defined in a much more general framework than for Free-Wilson analysis to be able to cope with very diverse structures. Such problems may arise much more frequently than is commonly believed. Typical cases are where data from large data bases, from mass screening or data collections from literature are to be analyzed.

Topological (Substructural) Descriptors

For a better understanding of the following, a definition of how the terms *feature* and *descriptor* will be used throughout this paper seems appropriate. We will call topological features any two-dimensional fragment of chemical compounds used to characterize their structure in a topological analysis. Topological descriptors then describe the occurrence of such features in each compound. In the case of Free-Wilson analysis, for example, a feature is a particular substituent in a defined position of substitution, and the corresponding descriptor is defined according to Eq. (1). There are many types of topological descriptors which all have in common that they can directly be derived either from two-dimensional chemical structures or from connection tables; excellent reviews are presented, for example, by Stuper et al. (*8*) and by Bawden (*18*). The simplest possible descriptors are counts of atoms or bonds of specified types which, however, contain very little information about the topology of molecules so that completely different molecules may have identical descriptor values. For this reason such descriptors will not be discussed here although they have occasionally been applied in QSAR analysis (*18*).

Part of the structural information lost in the simple atom and bond descriptors can be restored if more complex features comprising larger parts of chemical structures (substructures) are considered. There are several ways of doing that. First of all, features can be generated following certain rules algorithmically starting from single atoms or bonds as centers leading to atom-centered and bond-centered fragments. Such fragments are defined in terms of concentric areas of structure surrounding each atom (except hydrogens) or bond at different levels of complexity and specificity with information about atom and bond types, nonhydrogen connections, etc., in this area. Very commonly used are the so-called augmented atom fragments that describe atoms with their next neighbors. In deriving these fragments, each nonhydrogen nonterminal atom is once considered as a center. Augmented atom fragments may be further extended by enlarging the area considered. Ganglia-augmented atom fragments, for example, are obtained if the additional bonds on all atoms are added to an augmented atom fragment. The set of features generated depends on the structures present in the training series. Problems of redundancy may arise if features of different complexity are to be used in the same analysis.

Another possibility is to define a library of features (substructures) in such a way that basic topological characteristics are adequately represented. Such characteristics can be, for example, rings, functional groups, heteroatoms or other centers believed to be important for drug–biosystem interaction and distances between such centers (paths). Although linear notations as used for computer storage and retrieval of chemical structures, e.g., WLN (*19*), can serve as a source of fragment codes of this type, it seems advisable to select a more problem oriented library on the basis of "common sense" and the experience of the researcher (*20*). According to Kirschner and Kowalski (*21*) such a library should include features that "are thought to provide at least some information related to the biological activity, are known for the greatest majority of the compounds and if found to be related to activity, will provide the chemist with some degree of insight into the mechanism of action." To this end a basic library of potentially active centers such as atoms or groups of atoms (functional groups) likely to be involved in drug–receptor interaction via van der Waals or other forces may be set up and used to create the concrete features for each particular problem via a set of rules (a language) in an open-ended way. Cases in point are the SSFN and the DCAM systems that are based on the predefined "descriptor centers" and distances between them (*22–24*).

Even with more or less complex substructural features, a considerable part of structural information is still lost. This situation may be overcome in either of two ways. Following the approach of Jurs and colleagues (*8*) additional descriptors can be defined that characterize the environment of the substructures (environment descriptors). A better strategy is to code not only for the type of substructures but also for the molecular region where they occur whenever this is feasible. A typical example is Free-Wilson analysis which, however, is restricted to more or less congeneric series. For more heterogeneous data sets more general template models become necessary. A template can be generated by superimposing all structures of the starting set in such a way that all features of these structures considered as potentially important (descriptors centers) are unambiguously and adequately represented. As a result an artificial reference diagram is obtained that may be considered as a hypothetical parent of all compounds. Descriptors are then derived by comparing the structure of each compound with this template. Such approaches have, for example, been used by Cammarata and Menon (*25,26*) by Henry and Block (*27,28*), and in our laboratory (*15,29–32*); philosophically, they are similar with the "hyperstructure" of the DARC/PELCO system (*33*) and of the MTD approach of Simon and Colleagues (*34*).

Once the features are defined each compound is characterized by a set (a vector) of descriptor values (one for each feature). The simplest way to define such values is that according to Eq. (1), which is particularly suitable for complex template approaches and is also used in the methods LOGANA and LOCON to be discussed later. The descriptor values may also be defined as occurrence numbers of the features in the compounds according to

$$x_{ij} = \begin{cases} k, & \text{if feature } i \text{ occurs } k \text{ times in the } j\text{th compound} \\ 0, & \text{if feature } i \text{ is not present in the } j\text{th compound} \end{cases} \tag{3}$$

Finally, a third possibility is to derive descriptors on the basis of graph theory or pragmatic rules. An example of the first type is molecular connectivity (*35*), and of the latter type, are the environment descriptors of Jurs (*8*).

In certain types of analysis topological descriptors are used together with some physicochemical parameters as, for example, molar refractivity or log P. This can be advantageous in treating complex problems but has the danger that the interpretation of the results may become extremely difficult due to complicated relationships between the two descriptor sets. As will be shown later, physicochemical quantities can also be transformed into binary variables.

Using topological descriptors one is, of course, asking less of the data than with, for example, extrathermodynamic parameters, but such descriptors are the only possibility for large data sets of high structural diversity, and the results can still effectively be used for guiding synthesis and biological testing. The crucial step always is the selection of features. If the features are in error no meaningful results can be expected. In many cases this implies that compromises have to be made and that different types of features must be used in the same analysis. There are a number of problems that may arise when defining the features such as redundancy or ambiguity. These problems, however, are so special and complicated that they cannot be discussed here.

Principles of Topological Analysis

There are many possible ways to look for relations between topological descriptors and biological properties (classical QSAR methods where topology-derived descriptors such as indicator variables or molecular connectivity are used will not be considered here). There are usually two objectives: to recognize substructures "typical" of a particular biological effect and to predict for new compounds whether they will be (highly) active with respect to this effect. Which of the two is in the foreground depends on the data set, the descriptors and the procedure used. It may be said, however, that most of the procedures so far described in the literature perform better with respect to the latter; exceptions are the methods LOGANA and LOCON to be discussed in this paper which were especially designed for evaluating "topological pharmacophores."

From a methodological point of view three types of approaches may be differentiated:* heuristic "activity index" techniques; classification (pattern recognition) methods; and topological pattern finders. This differentiation is not very sharp (especially not between the first two categories) but helpful to systematize the various procedures.

Heuristic index techniques start from a classification of the compounds of the training series with respect to the biological activity of interest (usually, two classes, active versus inactive compounds, are used). The distribution of the selected features over the classes is then analyzed, and based on frequencies and/or probabilities of occurrence each feature is assigned an "activity index" which expresses how important it is for the active class. New compounds are then classified by summing the indices of their features. A typical approach of this kind is the substructural analysis introduced by Cramer et al. (*38*), which represents the first large scale computerized method for topological analysis. The compounds of a training series were divided into two classes (active and inactive) and fragmented into substructures using a library of atom, bond and substructure features. For each feature the so-called substructure activity frequency (SAF) is then calculated according to:

$$\mathrm{SAF}_i = \frac{\text{number of active compounds containing the feature } i}{\text{total number of compounds containing the feature } i} \qquad (4)$$

The SAF_i value embodies the probability contribution of the i-th feature to the overall probability that the compound containing that feature will be biologically active. To characterize the activity of compounds the mean substructure activity frequency (MSAF) is used. For the j-th compound this value is

$$\mathrm{MSAF}_j = \frac{1}{m_j} \sum_i x_{ij}\mathrm{SAF}_j \qquad (5)$$

where m_j is the number of features (fragments occurring in the j-th compound and x_{ij} a topological descriptor defined according to Eq. (2). As shown by an example concerning the immunoregulatory potency of 770 compounds, the MSAF values are indeed connected with biological potency so that the probability of a compound being active will be the higher the higher its MSAF value is. An unequivocal classification of compounds simply by means of MSAF values, however, is not possible since for that purpose a decision rule would be required.

Statistically more sophisticated is the approach of Hodes and colleagues (*39–42*) who devised a heuristic method for the automated selection of drugs for antitumor screening with the objective of increasing the incidence of biologically active compounds. Again, the compounds of a training series are divided into two classes (active/inactive) and described by a library of topological features (ring, nucleus, augmented atom and ganglion augmented atom fragments). Each fragment is then tested to determine whether its incidence in both classes is significantly different from the randomly expected incidence. A particular feature i will gain in importance for the class considered as this difference

*There are some applications where substructural features as described in the preceding section are used as variables in multiple regression analysis (*17,36,37*). This again implies at least in part that a linear additive model is supposed to hold which will be an exception rather than a rule. For this reason approaches of this type will not be discussed here.

becomes larger, and this may be expressed by the reciprocal of the probability P_i that this difference can be generated by chance. On this basis, an activity index $I_{i,a}$, and an inactivity index $I_{i,in}$ can be defined for each feature i.

$$I_{i,a} = 1/P_{i,a} \quad (6)$$

$$I_{i,in} = 1/P_{i,in} \quad (7)$$

For any given compound X with fragments $i = 1, \ldots, n$, an activity and an inactivity score can now be computed by simply multiplying the respective $I_{i,a}$ or $I_{i,in}$ values. On numerical grounds, it is desirable to perform a logarithmic transformation and to define these scores as

$$I_{X,a} = \sum_{i=1}^{n} \log I_{i,a} \quad (8)$$

$$I_{X,in} = \sum_{i=1}^{n} \log I_{i,in} \quad (9)$$

The scores provide a measure of the probabilities that compound X will belong to the respective classes. If potentially active compounds are to be selected for a screening program, both the activity as well as the inactivity score have to be taken into account, which can be done in various ways depending on the needs of the selection program. A very simple way is to combine both scores into a single quantity such that higher values establish high priorities for testing. An extended version of this procedure has successfully been applied to mutagenicity data by Tinker as a predictive test for hazard evaluation (*43,44*). Philosophically similar is the ORACLE procedure (*22,24*), which was developed for the purpose of guiding the screening of compounds in a battery of biological tests. The basis of ORACLE is a structure and pharmacological activity file including about 6000 compounds and 55 major types of pharmacological activities. The compounds of the data file are divided into 55 classes according to the types of their biological activity. Each class is then separately compared with the rest of the compounds contained in the remaining 54 classes in order to find descriptors (derived from SSFN; see above) that represent the particular type of activity exhibited by the class under investigation. A descriptor is considered to represent a particular activity feature if its presence in this class is not a chance event. Sets of descriptors are found in this way which are regarded to be typical of each respective class. Once these sets are known, the type of pharmacological activity to be expected for a new compound can be predicted so that it becomes possible to decide what compound should be investigated for which of the 55 pharmacological effects. The method will fail if a pharmacological effect can be produced via different mechanisms of action. Nevertheless, ORACLE has successfully been applied. The system correctly recognized the presence of earlier established activities for the majority of compounds in the data base, and a number of unknown activities could be predicted for several compounds.

Another program of the index type which, however, is already very close to classification techniques is the STRAC procedure designed for lead optimization (*22*). Again, the compounds of the starting set are divided into the classes "active" and "inactive" and described by topological features. These features are selected as in Free-Wilson analysis as substituents in their position of substitution so that STRAC is limited to more or less homologous series. The advantage over the Free-Wilson analysis is that no assumptions about the additivity of substituent effects are required. A feature is considered discriminating if the probability that a compound containing this feature belongs to one of the two classes is greater than a certain threshhold value. In many cases the discriminating features found in this way are not sufficient to classify all compounds of the training series. New discriminating features are then derived by logical operations in an interactive process until a sufficient separation of the two classes is achieved. The features so obtained are either typical for the active ("activity features") or for the inactive ("inactivity features") analogs. New compounds can then be classified according to their predominant features. A compound is rated active if

$$V_a - V_{in} > \eta \quad (10)$$

and inactive if

$$V_{in} - V_a > \eta \quad (11)$$

where V_a and V_{in} are the occurrence numbers of activity and inactivity features, respectively, in this compound and η is a certain threshold value.

Methods of the second category, classification techniques, also start from a classification of the compounds of the training series but use different strategies as the index approaches. There is a variety of methods available which will not be discussed here in detail (*8,11,15,21*). The principle usually is to find a classifier that is a mathematical expression containing the descriptors as variables. In most (but not all) cases such classifiers are weighted linear combinations of variables that are found to make a significant contribution to class separation. With the help of the classifiers and decision rules that can be based on statistical or geometrical (distances in parameter space) criteria, the activity of new compounds can be predicted. The methods of choice for topological descriptors are the so-called "nonprobabilistic" or "non-parametric" techniques such as, for example, the linear learning machine or the k-nearest-neighbors method, since these methods do not require certain statistical data distributions that are not likely to be fulfilled for such variables. Both index as well as classification methods make the implicit assumption that the contribution of a given substructural unit to the biological activity is a consistent factor. This, however, is frequently not true where large and structurally diverse data sets for relatively unspecified biological effects are considered, since in such cases the biological mechanism of action is likely to be different for different subsets of compounds. In addition, complex intramolecular interactions not accounted for by simple descriptors may strongly influence the role of certain substructures; in the extreme case they may be favorable for activity in one particular chemical environment and unfavorable in another. In such a situation the division of compounds into classes with the sole criterion of measured biological activity values is basically wrong, since the classes so obtained do not constitute homogeneous and well-defined entities but consist of different clusters with quite different characteristics. None of the methods mentioned so far take this fact into account. This has two serious consequences that one must be well aware of: (1) a certain percentage of predictions is bound to be wrong; (2) a classifier obtained from such data may become very complex since it actually presents a mixture and a compromise of several classifiers needed to separate different subsets of compounds in the active class from the inactive compounds. Such classifiers admit no chemical interpretation, and even though they may still be used as predictive tools in a purely algorithmic sense (keeping in mind point *1*), this situation is far from being satisfactory for a medicinal chemist or toxicologist.

Some of the problems mentioned above can be avoided by the third category of topological methods, which we have termed in this paper "topological pattern finders." Let us first define what we understand by a "topological pattern" in this context. We regard it as an ensemble of substructural features that is characteristic of a group of compounds possessing a desired biological property (e.g., high activity) but absent from compounds devoid of this property. This ensemble, which may be called a "topological pharmacophore," is always considered in its entirety. This means, in particular, that no conclusions about the importance of isolated features on biological activity is always seen to be dependent on the other features present in the pharmacophore. In this way the assumption that the contribution of a given structural unit to the biological activity is a consistent factor is no longer necessary; as discussed above this assumption probably is not realistic for too complex problem. Furthermore, it becomes possible to handle data sets where the compounds do not act via a uniform mechanism. If different mechanisms operate they will be reflected by different topological pharmacophores. It was for these reasons that we started to develop topological techniques of the "pattern finding" type (*15,29–32,45*). Two of them, the methods LOGANA and LOCON (*45*), will be discussed in the next sections in some detail together with a simple example of application.

The LOGANA and LOCON Methods

In contrast to the index and classification methods, where activity indices or weights are assigned to single features, the basic structure of LOGANA and LOCON is the stepwise and interactive construction of combinations of such features using the logical operations "and," "or," and "not." These combinations are evaluated in such a way that they are typical of compounds possessing the biological property of interest to the highest extent (e.g., compounds with high or very high activity, compounds devoid of toxicity, etc.). They represent more or less complex structural patterns that may be regarded as "topological pharmacophores" and that can then aid in the design of new compounds. The philosophy of both methods is to make use of the data handling capacity of computers while maintaining an optimal impact of the researcher's professional skill and experience with no requirement for any special mathematical knowledge. The computer is used only to digest large data sets and condense the information inherent in them so that it becomes manageable. While this part is fully formalized, the real decision-making and all the conclusions to be drawn for further synthesis or testing are completely left to the researcher. This implies that neither assumptions regarding probabilities or data distribution nor mathematical models are necessary.

The program LOGANA starts from a classification of the compounds of the training series according to a biological activity score which allows the analysis of more or less crude data as from biological mass screening or sampled from different sources. LOCON, on the other hand, was designed to deal with continuous biological activity data in order not to lose information in those cases where the data are precise enough to allow for comparisons on such a scale. Both methods are binary descriptors as defined by Eq. (1), and they will be most efficient for template model derived features. Molecular properties like, for instance, hydrophobicity can also be included to a certain extent. This is done by selecting region(s) (or threshholds) of the corresponding molecular parameter (e.g., the log P or π) and defining $x_{ij} = 1$ if the jth compound falls into the ith region and $x_{ij} = 0$, if not.

The set of descriptors obtained according to Eq. (1) or the above definition can be extended by the logical operations "not" (negation, symbolized by a "~"sign) and "or" (disjunction, symbolized by a "$\vee$" sign). In a negation, the definition presented in Eq. (1) is reversed, meaning that the absence of the ith feature is considered important and used as a new feature:

$$x_{ij} = \begin{cases} 1, & \text{if feature } i \text{ is not present in the } j\text{th compound} \\ 0, & \text{if it is present} \end{cases} \quad (12)$$

A disjunction combines several features which are regarded to be so similar that they can be exchanged without noticeable effect on biological activity ("bioisosteric" features) into a new one. A disjunction of two descriptors x_i and x_k, for example, would read:

$$x_{ij} \vee x_{kj} = \begin{cases} 1, & \text{if features } i \text{ or } k \text{ are present in the } k\text{th compound} \\ 0, & \text{if neither feature } i \text{ nor feature } k \text{ is present} \end{cases} \quad (13)$$

LOGANA

Input into the program are the variables (topological descriptors) and a classification with respect to the biological activity of interest (e.g., high activity) into two classes (class 1: compounds with the desired biological property; class 2: desired property not present) for all compounds of the training series. Searched for are combinations of features (conjunctions, see below) that are present in as large as possible groups of class 1 compounds and absent or nearly absent in class 2 compounds. It is this strategy that makes LOGANA basically different from classification methods since these methods aim at a complete separation of the two classes. By considering subgroups it becomes possible to account for different mechanisms of action which may require different pharmacophores. Another point is that a limited set of not too simple topological features by its very nature frequently implies the formation of subgroups. Take, for example, the case that for a certain type of activity a hydrogen donor is needed in some particular region of the molecules while the chemical nature of this donor is not important and that different donor types appear in the compounds of the training series, each described by a different variable. In this case different topological patterns will result, each representing one subgroup of compounds according to the donor type. Only if the different donor descriptors are combined into one new variable by the logical "or" [disjunction, see Eq. (13)] will the division into subgroups disappear but this would require that a corresponding decision is made by the researcher.

The construction of more complex from the simple features is based on the logical operation "and" (conjunction, symbolized by a "$\wedge$" sign). A conjunction of two variables x_i and x_k is defined as

$$x_{ij} \wedge x_{ik} = \begin{cases} 1, & \text{if features } i \text{ and } k \text{ are present in the } j\text{th compound} \\ 0, & \text{if not} \end{cases} \quad (14)$$

Obviously, a conjunction represents a more complex chemical entity than the single variables and will, as a consequence, usually be present in a smaller set of compounds. A set of compounds having the structural features expressed by a conjunction in common will be called an "object group." Those conjunctions can be regarded best which yield object groups containing high numbers of class 1 compounds and the smallest possible number of class 2 compounds. As a measure of this property a simple quality criterion T can be formulated as

$$T = \frac{N_1 + (n_1 - n_2)}{N_1 + N_2} \quad (15)$$

with T normalized to

$$0 \leqslant T \leqslant 1$$

where N_1 = number of compounds in class 1, N_2 = number of compounds in class 2, n_1 = number of class 1 compounds in the object group, and n_2 = number of class 2 compounds in the object group.

The evaluation of the conjunction is preferred in a stepwise procedure where one more variable is added in each step using T as a selection criterion. In the first step all variables (including negations) are arranged in the order of descending T values and the best m variables (m adjustable) are selected as a starting set for the next step. Conjunctions of each of the variables of this starting set with all other variables one at a time are then formed according to Eq. (14) and the best m of these conjunctions are then printed out and transferred to the next step as a new starting set. Each conjunction of this set is then again combined with all variables (one at a time) by the logical operation "and," and the best m of the resulting conjunctions now comprising three variables are again printed out and transferred to the next step. The procedure is continued until a preset number of steps has been performed. This step-up procedure can be combined with feature elimination steps at each level. Applied to the m best conjunction comprising k variables after the kth forward step, backward elimination will once eliminate each variable from each conjunction and yield m "best" new conjunctions now comprising $k - 1$ variables. The main purpose of this option is to check for consistency.

Disjunctions [see Eq. (13)] will not be automatically formed in order to limit the resulting conjunctions to a manageable number. They can be introduced by a special option at any stage of the procedure. Along with each conjunction the following information appears in the printout: T, n_1, n_2; the compounds of the object group identified by the respective row numbers of the input data matrix; and accompanying variables. Accompanying variables are those that can be added to a conjunction without eliminating compounds from the object group. They must, of course, be considered when evaluating the results since they characterize structural features also present in all compounds of the object group.

The best conjunctions obtained after the analysis is completed can directly be retranslated into topological structures. In doing that and when interpreting the results the whole picture including the development of the conjunctions, the object groups described by them and the accompanying variables must be viewed. It is important to stress that the quality criterion T is not to be used for decision-making by the researcher but is only an internal tool of the program. For a detailed analysis it is usually advisable to investigate several classifications in parallel as, for example, all active versus inactive, highly active versus weakly active and/or inactive compounds, etc.

For linguistic simplicity it is tempting and practical to label features appearing in the best conjunction as"favorable" (or, in the case of negations, as "unfavorable") for the biological activity considered. Strictly speaking, this is not correct. All that can be concluded is that features represented in the best conjunctions are typical of the class 1 compounds, and that only in the context of these conjunctions and insofar, as the training series is representative of the chemical compound space.

Because of the formation of subgroups, LOGANA is not a procedure suitable for feature selection in order to find a classifier. A special option of the program, however, allows the evaluation of alternatives which usually are good classifiers. Alternatively, several conjunctions are combined by a logical "or" in such a way that n_1 is maximized under the condition of an optimal separation of class 1 and class 2 compounds. The alternative presents those conjunctions that optimally supplement each other and provides information concerning those objects of class 1 that can only be characterized together with certain objects of class 2.

LOCON

LOCON searches for conjunctions expressing such topological patterns that the corresponding object groups (compounds having these patterns in common) show an as high as possible mean value of biological activity. In other words, topological patterns are searched for which are characteristic of the most active compounds. Input into the program are the variables (topological descriptors) and biological activity values for all compounds of the training series. Biological activity is expressed here on a continuous scale using quantities such as, for example, ED_X, I_X, or LD_X, etc. Again, conjunctions are derived in a stepwise procedure by the logical operation "and " [Eq. (14)]. The quality criterion used here to characterize conjunctions (or variables) is defined as

$$D = (\mathrm{MS} - \mathrm{MO})\left[1 + \frac{(\mathrm{NS})(W)}{(\mathrm{NO})}\right] \qquad (16)$$

where MS = biological activity mean within the object group described by the respective conjunction; MO = overall mean (whole training series); NS = number of compounds in the object group; NO = number of compounds in the training series; W = adjustable weight (≥ 0).

The higher the values of D become the more characteristic is the corresponding conjunction of the high activity compounds. With the weight W the influence of the size of the object group on D can be adjusted. If this is regarded to be of secondary importance W is set equal to zero, and D will then solely depend on how much MS exceeds MO. As the T value in LOGANA, D is only an internal tool of the program and not to be used for interpretation and evaluation of the results.

In the first step of the analysis the variables are arranged in the order of descending D values, and the best variables are selected as starting set for the next step. Conjunctions of these best variables with all other variables (including negations) one at a time are then formed, and for each variable the m best (m adjustable) conjunctions as judged by D are printed out together with D, MS, the object group (compounds possessing the structural features expressed by the conjunction), the standard deviation, s, of biological activity in the object group, outliers (compounds with a biological activity differing more than ks from MS; k adjustable), and accompanying variables (see above). From these conjunctions the operator now selects those which he considers the most promising or meaningful according to all the available information and to his chemical intuition by inspection. These are then used as starting conjunctions and combined with all variables (one at a time) in the next step. The best m of the resulting new conjunctions now comprising three variables are again printed out for each conjunction of the starting set. A new starting set is then selected by inspection and extended by one variable in the next step. The procedure is continued until the "quality" of the resulting conjunction(s) as judged by D, MS, S and NS cannot be further improved.

Concerning the interpretation of the results all that has been said for LOGANA is also valid for LOCON.

Example: Carboxamides as Inhibitors of Succinate Dehydrogenase

As a test case, LOGANA and LOCON (*30*) were applied to data on the inhibition of succinate dehydrogen-

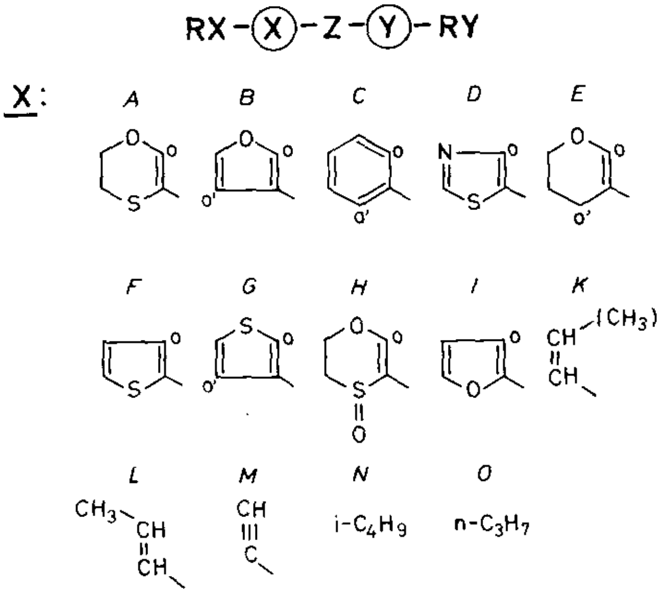

FIGURE 1. General structure of the carboxamides.

Table 1. pI_{50} values for the inhibition of succinate dehydrogenase from *Cryptococcus laurentii* by carboxanilides of the general structure depicted in Figure 1.[a]

No.	Z	X	$(RX)_{o,o'}$	$(RX)_{m,p}$	Y	RY	pI_{50}
01	CONH	A	CH_3	4H	Phenyl	3′-CH_3	4.00
02	CONH	B	CH_3, H	CH_3	Phenyl	3′-CH_3	3.46
03	CONH	A	CH_3	4H	Phenyl	2′,3′-CH_3	3.30
04	CONH	C	C_2H_5, H	3H	Phenyl	3′-CH_3	3.30
05	CONH	A	CH_3	4H	Phenyl		3.26
06	CONH	A	CH_3	4H	Phenyl	2′-C_6H_5	3.26
07	CONH	D	CH_3	NH_2	Phenyl		3.22
08	CONH	D	CH_3	NH_2	Phenyl	3!-CH_3	3.22
09	CONH	D	CH_3	NH_2	Phenyl	3′-OCH_3	3.22
10	CONH	B	CH_3, H	CH_3	Phenyl	3′-Cl	3.14
11	CONH	A	CH_3	4H	Phenyl	2′-CH_3	3.12
12	CONH	A	CH_3	4H	Phenyl	4′-Cl	3.00
13	CONH	E	C_2H_5, 2H	4H	Phenyl		3.00
14	CONH	C	I, H	3H	Phenyl		3.00
15	CONH	B	CH_3, H	CH_3	Phenyl	2′-CH_3	2.72
16	CONH	A	CH_3	4H	Phenyl	2′,4′-CH_3	2.70
17	CONH	A	CH_3	4H	Phenyl	2′-C_2H_5	2.70
18	CONH	A	CH_3	4H	Phenyl	2′-Cl	2.70
19	CONH	E	CH_3, 2H	4H	Phenyl	3′-OCH_2O-4′	2.70
20	CONH	B	CH_3, H	CH_3	Phenyl		2.66
21	CONH	A	CH_3	4H	Phenyl	2′-OCH_3	2.60
22	CONH	E	CH_3, 2H	4H	Phenyl	3′-OCH_3	2.60
23	CONH	C	C_2H_5, H	3H	Phenyl	2′,3′-CH_3	2.60
24	CONH	C	C_2H_5, H	3H	Phenyl		2.57
25	CONH	D	CH_3	CH_3	Phenyl		2.52
26	CONH	F	CH_3	2H	Phenyl		2.48
27	CONH	B	CH_3, H	CH_3	Phenyl	2′,3′-Cl	2.46
28	CONH	F	CH_3	2H	Phenyl	2′,3′-CH_3	2.46
29	CONH	C	CH_3, H	3H	Phenyl	3′-CH_3	2.41
30	CONH	A	CH_3	4H	Phenyl	2′-CH_3, 4′-OCH_3	2.40
31	CONH	A	CH_3	4H	Phenyl	2′-CH_3, 4′-Cl	2.40
32	CONH	C	Br, H	3H	Phenyl		2.40
33	CONH	B	CH_3, H	CH_3	Phenyl	4′-Br	2.35
34	CONH	C	CH_3, H	3H	Phenyl	3′-OCH_3	2.26
35	CONH	A	CH_3	4H	C_6H_{11}		2.25
36	CONH	E	CH_3, 2H	4H	Phenyl		2.22
37	CONH	B	CH_3, H	CH_3	C_6H_{11}		2.22
38	CONH	G	CH_3, H	CH_3	Phenyl	2′,3′-CH_3	2.19
39	CONH	G	CH_3, H	CH_3	Phenyl	2′-CH_3	2.12
40	CONH	G	CH_3, H	CH_3	Phenyl		2.11
41	CONH	C	CH_3, H	3H	Phenyl	2′,3′-CH_3	2.00
42	CONH	B	CH_3, H	H	Phenyl		2.00
43	CONH	F	CH_3	2H	Phenyl	2′-CH_3	1.98
44	$CONCH_3$	A	CH_3, H	4H	Phenyl		1.96
45	CONH	D	CH_3	CH_3	Phenyl	2′-CH_3	1.96
46	CONH	A	CH_3, H	4H	Phenyl	2′,6′-CH_3	1.89
47	CONH	C	C_2H_5, H	3H	Phenyl	2′-CH_3	1.87
48	CONH	A	CH_3, H	4H	Phenyl	2′,6′-C_2H_5	1.85
49	CONH	F	CH_3	2H	Phenyl	3′-CH_3	1.84
50	CONH	C	CH_3, H	3H	Phenyl		1.74
51	CONH	C	CH_3, H	3H	Phenyl		1.58
52	CONH	C	OH, H	3H	Phenyl	3′-CH_3	1.48
53	CONH	G	CH_3, H	CH_3	Phenyl	2′-CH_3	1.42
54	CONH	C	OH, H	3H	Phenyl	2′-CH_3	1.40
55	CONH	F	CH_3	2H	C_6H_{11}		1.22
56	CONH	C	CH_3, H	3H	Phenyl	2′-CH_3	1.08
57	CONH	C	OH, H	3H	Phenyl		1.07
58	CONH	A	CH_3, H	4H	Phenyl	3′,4′-Cl	1.00
59	CONH	H	CH_3, H	4H	Phenyl		1.00
60	CONH	G	CH_3, H	CH_3	C_6H_{11}		0.98
61	$CONCH_3$	E	CH_3, 2H	4H	Phenyl		0.92
62	CONH	C	CH_3, H	3H	Phenyl	2′-OCH_3	0.74
63	CONH	I	CH_3	2H	Phenyl		0.60
64	CONH	F	H	CH_3, H	Phenyl	2′-CH_3	0.42
65	CONH	C	C_2H_5, H	3H	C_4H_9		0.40
66	CONH	K	CH_3		Phenyl		0.35
67	CONH	F	H	CH_3, H	Phenyl	3′-CH_3	1.0
68	CONH	C	H, H	2H, CH_3	Phenyl		< 0.6
69	CONH	A	CH_3, H	4H	H		< 0.3

(continued)

Table 1 (Continued)

No.	Z	X	$(RX)_{o,o'}$	$(RX)_{m,p}$	Y	RY	pI_{50}
70	COOH	A	CH_3, H	4H	—		< 0.3
71	CONH	E	C_6H_5, 2H	4H	Phenyl		< 0.3
72	CONH	C	H, H	3H	Phenyl		< 0.3
73	CONH	C	F, H	3H	Phenyl		< 0.3
74	CONH	C	H, H	CH_3, 2H	Phenyl		< 0.3
75	$COCH_2$	B	CH_3, H	H	Phenyl		< 0.3
76	CONH	I	H	CH_3, H	Phenyl		< 0.3
77	CONH	F	CH_3	2H	CH_3		< 0.3
78	CONH	F	CH_3	2H	C_4H_9		< 0.3
79	CONH	F	H	CH_3, H	Phenyl		< 0.3
80	CONH	F	H	CH_3, H	Phenyl	2′,3′-CH_3	< 0.3
81	CONH	F	H	CH_3, H	C_6H_{11}		< 0.3
82	CONH	O			Phenyl		< 0.3
83	CSNH	C	CH_3, H	3H	Phenyl		< 0.1
84	SO_2NH	C	CH_3, H	3H	Phenyl		< 0.1
85	CSNH	B	CH_3, H	H	Phenyl		0.0
86	CONH	E	H, 2H	4H	Phenyl		0.0
87	CONH	M			Phenyl		0.0
88	CONH	L			Phenyl		0.0
89	CONH	N			Phenyl		0.0

[a] Data from White and Thorn (*46*).

ase by a series of antifungal carboxamides (*46*) of the general structure presented in Figure 1. This example was selected because the training series shows sufficient structural diversity and the data allow the application of both, LOGANA and LOCON (the former after division of the compounds into classes; see below).

All compounds and their activities are summarized in Table 1. The features used in this case are selected so that not only the type of substructures but also the region in the molecules where they are situated can be coded for. Table 2 represents the features used in LOGANA and Table 3 some additional features applied in LOCON only. These additional features were defined after the results of LOGANA were known with the intention to increase the sharpness of the LOCON analysis. We have always found it useful to apply LOGANA even to continuous data (after introducing a classification) prior to a LOCON analysis since LOGANA works faster and its results may be used to improve or complete the feature space for the more powerful LOCON procedure. Finally, Table 4 contains the code of all compounds in terms of the binary descriptors defined according to Eq. (1).

LOGANA Results. LOGANA was applied to the following classifications:

Problem A: active versus inactive compounds

Class 1(A) $0.60 \leqslant pI_{50} \leqslant 4.00$ $N_{1(A)} = 64$

Class 2(A) $pI_{50} < 0.60$ $N_{2(A)} = 25$

Problem B: highly versus weakly active compounds

Class 1(B) $2.00 \leqslant pI_{50} \leqslant 4.00$ $N_{1(B)} = 40$

Class 2(B) $0.60 \leqslant pI_{50} \leqslant 1.70$ $N_{2(B)} = 19$

Problem C: very active versus inactive compounds

Class 1(C) $3.00 \leqslant pI_{50} \leqslant 4.00$ $N_{1(C)} = 14$

Class 2(C) $pI_{50} < 0.60$ $N_{2(C)} = 25$

The development of the best conjunction obtained for problem A [conjunction (I)] is presented in Table 5*. This conjunction is present in almost all class 1(A) compounds and does not occur in any of the inactive analogs [class 2(A) compounds]. It can directly be translated into the structure shown in Figure 2. This structure may be regarded to present basic structural requirements for activity which include an intact amide group (variable $Z1$), a substituent different from H, F or phenyl in the *ortho* position of X (variables $RX7$, $RX8$, $RX9$), a C = C group adjacent to the carbon atom of the amide moiety which is part of a ring, and a ring adjacent to the amide nitrogen. Another conjunction of interest for a subset of class 1(A) compounds is:

$$X16 \wedge \sim Y6 \wedge (Z1 \vee Z2) \wedge RX1 \qquad \text{(II)}$$

with $n_{1(A)} = 30$, $n_{2(A)} = 0$

The compounds of this subset have an oxygen adjacent to the C = C group in Figure 1 (feature $X16$) and methyl in $(RX)_{ortho}$. It cannot be said whether this more specified pattern provides higher activity than that in Figure 2 but it is certainly of interest to know that it is present in about 50% of the class 1(A) compounds while also absent from all inactive analogs.

Some additional information about what features may be important for gradation of activity within the active compounds was expected from problem B. The best conjunction obtained for that problem is

$$RY16 \wedge \sim X6 \wedge \sim RY14 \wedge \sim RY15 \wedge \sim X7 \wedge RX1 \qquad \text{(III)}$$

with $n_{1(B)} = 21$, $n_{2(B)} = 2$

Accompanying variables: $Z1$, $Y1$, $X8$, $X9$, $RX11$

Table 2. Definition of descriptors x_i used in LOGANA: x_i = 1 if feature i is present.[a]

Feature	Definition	
Z1	-CONH-	in Z
Z2	$-CONCH_3-$	in Z
Z3	-COOH	in Z
Z4	$-COCH_2-$	in Z
Z5	-CSNH-	in Z
Z6	$-SO_2NH-$	in Z
X1 - X14	X = structures A - O	
X15	- C = C - C = O	in X
X16	- O - C = C - C = O	in X
X17	- S - C - C = O	in X
RX1	$-CH_3$	in $(RX)_o$
RX2	$-C_2H_5$	in $(RX)_o$
RX3	-I	in $(RX)_o$
RX4	-Br	in $(RX)_o$
RX5	-Cl	in $(RX)_o$
RX6	-OH	in $(RX)_o$
RX7	-H	in $(RX)_o$
RX8	-phenyl	in $(RX)_o$
RX9	-F	in $(RX)_o$
RX10	$-NH_2$	in $(RX)_o$
RX11	$-CH_3$	in $(RX)_{o,m,p}$
Y1	-phenyl	in Y
Y2	-cyclohexyl	in Y
Y3	-butyl	in Y
Y4	-H	in Y
Y5	-methyl	in Y
Y6	no ring	in Y
RY1	$3'-CH_3$	
RY2	$2'-CH_3$	
RY3	$2'-C_6H_5$	
RY4	$3'3'-OCH_3$	
RY5	3'-Cl	
RY6	4'-Cl	
RY7	$4'-CH_3$	
RY8	$2'-C_2H_5$	
RY9	2'-Cl	
RY10	$3'-OCH_2O-4'$	
RY11	$2'-OCH_3$	
RY12	$4'-OCH_3$	
RY13	4'-Br	
RY14	$6'-CH_3$	
RY15	$6'-C_2H_5$	
RY16	R_y with $\pi > 0$	

[a] Symbols as in Figure 1.

The structure obtained from this conjunction (Fig. 3) is present in about 50% of the very active and in only two out of the 19 weakly active compounds. The additional information as compared to conjunction (I) and (II) is that, for high activity, the ring attached to the nitrogen of the amide group should be phenyl (or, maybe, in a more general sense simply aromatic) and that hydrophobic substituents at this ring are favorable as long as an o, o'-disubstitution is avoided. Several other conjunctions also containing these but in addition some special X-ring features describe the rest of the class 1(B) compounds without, however, providing much useful additional information. For this reason they will not be discussed.

Table 3. Additional features derived for the application of LOCON: x_i = 1 if feature i is present.

X18	X10 $\vee$ X11 $\vee$ X12 $\vee$ X13 $\vee$ X14
X19	X2 $\wedge$ RX11
X20	X4 $\wedge$ RX10
X21	X4 $\wedge$ RX11
X22	X6 $\wedge$ RX11
X23	X7 $\wedge$ RX11
X24	X9 $\wedge$ RX11
RX12	RX7 $\vee$ RX9
RX13	RX3 $\vee$ RX4 $\vee$ RX5 $\vee$ RX9
RY17	RY2 $\vee$ RY3 $\vee$ RY8 $\vee$ RY11
RY18	RY14 $\vee$ RY15
RY19	RY6 $\vee$ RY7 $\vee$ RY12 $\vee$ RY13

Problem C, finally, yields the following results:

$$RY16 \wedge \sim X6 \qquad \text{(IV)}$$

with $n_{1(C)} = 10$, $n_{2(C)} = 0$

Accompanying variables: $Z1$, $\sim X5$, $\sim X9$, $\sim X10$, $\sim X11$, $\sim X12$, $\sim X13$, $\sim X14$, $\sim X15$, $\sim RX3$, $\sim RX7$, $\sim RX8$, $\sim RX9$, $Y1$

Conjunction IV leads to the structure presented in Figure 4 which shows all the basic features of conjunctions (I) and (II) but also provides some of the additional information obtained from conjunction (III) (see Fig. 3). The lesson to be learned from this is that compounds of medium activity may well be eliminated from a LOGANA analysis. As in the present and other examples [(*31*) and unpublished work] the result is usually sharper and by no means less representative as for the whole data set [conjunctions (I) and (II)]. Especially for very large training series a lot of work can be saved in this way.

LOCON Results. Two relatively small conjunctions comprising three variables already produce object groups

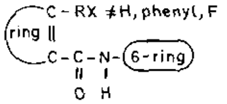

FIGURE 2. Structural pattern resulting from conjunction I.

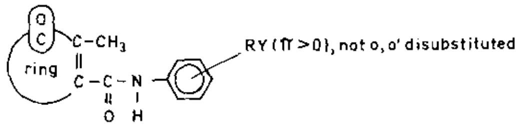

FIGURE 3. Structural pattern resulting from conjunction III.

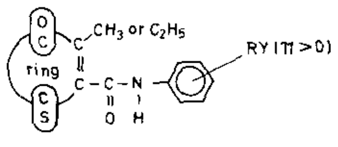

FIGURE 4. Structural pattern resulting from conjunction IV.

Table 4. Features having the value of $x_i = 1$ for all compounds.

Compound No.	Features used in both LOGANA and LOCON	Additional features used in LOCON only
1	Z1, X1, X15, X16, X17, RX1, Y1, RY1, RY16	RY20
2	Z1, X2, X15, X16, RX1, RX11, Y1, RY1, RY16	X19, RY17, Ry20
3	Z1, X1, X15, X16, X17, RX1, RY1, RY2	RY20
4	Z1, X3, X15, RX2, Y1, RY1, RY16	RY20
5	Z1, X1, X15, X16, X17, RX1, Y1	
6	Z1, X1, X15, X16, X17, RX1, Y1, RY3, RY16	RY17
7	Z1, X4, X15, X17, RX1, RX10, Y1	X20
8	Z1, X4, X15, X17, RX1, RX10, Y1, RY1, RY16	RY20
9	Z1, X4, X15, X17, RX1, RX10, Y1, RY4, RY16	X20, RY20
10	Z1, X2, X15, X16, RX1, RX11, Y1, RY5, RY16	X19, RY20
11	Z1, X1, X15, X16, X17, RX1, Y1, RY2, RY16	RY17
12	Z1, X1, X15, X16, X17, RX1, Y1, RY6, RY16	RY19
13	Z1, X5, X15, X16, RX2, Y1	
14	Z1, X3, X15, RX3, Y1	RX13
15	Z1, X2, X15, X16, RX1, RX11, Y1, RY2, RY16	X19, RY17
16	Z1, X1, X15, X16, X17, RX1, Y1, RY2, RY7, RY16	RY17, RY19
17	Z1, X1, X15, X16, X17, RX1, Y1, RY8, RY16	RY17
18	Z1, X1, X15, X16, X17, RX1, Y1, RY9, RY16	RY17
19	Z1, X5, X15, X16, RX1, Y1, RY10, RY16	RY19, RY20
20	Z1, X2, X15, X16, RX1, RX11, Y1	X19
21	Z1, X1, X15, X16, X17, RX1, Y1, RY11	RY17
22	Z1, X5, X15, X16, RX1, Y1, RY4, RY16	RY20
23	Z1, X3, X15, RX2, Y1, RY1, RY2, RY16	RY17, RY20
24	Z1, X3, X15, RX2, Y1	
25	Z1, X4, X15, RX1, RX11, Y1	X21
26	Z1, X6, X15, X17, RX1, Y1	
27	Z1, X2, X15, X16, RX1, RX11, Y1, RY5, RY9	X19, RY17, RY20
28	Z1, X6, X15, X17, RX1, Y1, RY1, RY2, RY16	RY17, RY20
29	Z1, X3, X15, RX1, Y1, RY1, RY16	RY20
30	Z1, X1, X15, X16, X17, RX1, Y1, RY2, RY12, RY16	RY17, RY19
31	Z1, X1, X15, X16, X17, RX1, Y1, RY2, RY6, RY16	RY17, RY19
32	Z1, X3, X15, RX4, Y1	RX13
33	Z1, X2, X15, X16, RX1, RX11, Y1, RY13, RY16	X19, RY19
34	Z1, X3, X15, RX1, Y1, RY4, RY16	RY20
35	Z1, X1, X15, X15, X17, RX1, Y2	
36	Z1, X5, X15, X16, RX1, Y1	
37	Z1, X2, X15, X16, RX1, RX11, Y2	X19
38	Z1, X7, X15, RX1, RX11, Y1, RY1, RY2, RY16	X24, RY17, RY20
39	Z1, X7, X15, RX1, X11, Y1, RY2, RY16	RY17
40	Z1, X7, X15, RX1, RX11, Y1	
41	Z1, X3, X15, RX1, Y1, RY1, RY2, RY16	RY17, RY20
42	Z1, X2, X15, X16, RX1, Y1	
43	Z1, X6, X15, X17, RX1, Y1, RY2, RY16	RY17
44	Z2, X1, X15, X16, X17, RX1, Y1	
45	Z1, X4, X15, X17, RX1, RX11, Y1, RY2, RY16	X21, RY17
46	Z1, X1, X15, X16, X17, RX1, Y1, RY2, RY14, RY16	RY17, RY18
47	Z1, X3, X15, RX2, Y1, RY2, RY16	RY17
48	Z1, X1, X15, X16, X17, RX1, Y1, RY3, RY15, RY16	RY17, RY18
49	Z1, X6, X15, RX1, Y1, RY1, RY16	RY20
50	Z1, X3, X15, RX1, Y1	
51	Z1, X1, X15, RX5, Y1	RX13
52	Z1, X3, X15, RX6, Y1, RY1, RY16	RY20
53	Z1, X7, X15, RX1, RX11, Y1, RY2, RY16	X24, RY17
54	Z1, X3, X15, RX6, Y1, RY2, RY16	RY17
55	Z1, X6, X15, X17, RX1, Y2	
56	Z1, X3, X15, RX1, Y1, RY2, RY16	RY17
57	Z1, X3, X15, RX6, Y1	
58	Z1, X1, X15, X16, X17, RX1, Y1, RY5, RY6, RY16	RY19, RY20
59	Z1, X8, X15, X16, X17, RX1, Y1	
60	Z1, X7, X15, RX1, RX11, Y2	X23
61	Z2, X5, X15, X16, RX1, Y1	
62	Z1, X3, X15, RX1, Y1, RY11	RY17
63	Z1, X9, X15, RX1, Y1	
64	Z1, Xt, X15, RX7, RX11, Y1, RY2, RY16	X22, RX12, RY17
65	Z1, X3, X15, RX2, Y3, Y6	
66	Z1, X10, X15, RX1, Y1	X18
67	Z1, X6, X15, RX7, RX11, Y1, RY1, RY16	RX12, RY20

Table 4. (Continued)

Compound No.	Features used in both LOGANA and LOCON	Additional features used in LOCON only
68	Z1, X3, X15, RX7, RX11, Y1	RX12
69	Z1, X1, X15, X16, X17, RX1, Y4, Y6	
70	Z3, X1, X15, X16, X17, RX1, Y6	
71	Z1, X5, X15, X16, RX2, RX8, Y1	
72	Z1⅛, X3, RX7, Y1	RX12
73	Z1, X3, RX9, Y1	RX12, RX13
74	Z1, X3, RX7, RX11, Y1	RX12
75	Z4, X2, X15, X16, RX1, Y1	
76	Z1, X9, X15, RX7, RX11, Y1	
77	Z1, X6, X15, X17, RX1, Y5, Y6	
78	Z1, X6, X15, X17, RX1, Y3, Y6	
79	Z1, X6, X15, X17, RX7, RX11, Y1	RX12
80	Z1, X6, X15, X17, RX7, RX11, Y1, RY1, RY2, RY16	RX12, RY17, RY20
81	Z1, X6, X15, X17, RX7, RX11, Y2	RX12
82	Z1, X14, Y1	X18
83	Z5, X3, X15, RX1, Y1	
84	Z6, X3, X15, RX1, Y1	
85	Z5, X2, X15, X16, RX1, Y1	
86	Z1, X5, X15, X16, RX7, Y1	
87	Z1, X13, Y1	X18
88	Z1, X2, X15, Y1	X18
89	Z1, X12, Y1	X18

Table 5. Development of the best LOGANA conjunction (no. I) obtained for problem A.

Conjunction	^{n}1A	^{n}2A
X15	64	19
X15 ∧ ~ RX7	63	12
X15 ∧ ~ RX7 ∧ ~ Y6	63	7
X15 ∧ ~ RX7 ∧ ~ Y6 ∧ Z1	61	3
X15 ∧ ~ RX7 ∧ ~ Y6 ∧ Z1 ∧ ~ (X10 ∨ X11)	61	1
X15 ∧ ~ RX7 ∧ ~ Y6 ∧ Z1 ∧ ~ (X10 ∨ X11) ∧ ~ RX8[a]	61	0

[a] Accompanying variables: X12, X13, X14, RX9.

Table 6. Development of a conjunction (no. VIII) describing a group of very active compounds.

Conjunction	MS	NS	s
X16 ∨ X20	2.21	40	1.099
(X16 ∨ X20) ∧ RY16	2.57	22	0.660
(X16 ∨ X20) ∧ RY16 ∧ ~ RY18	2.67	20	0.608
(X16 ∨ X20) ∧ RY16 ∧ ~ RY18 ∧ ~ RY19[a]	3.04	14	0.3666

[a] Accompanying variables: Z1, Y1, ~RX8, ~X3, ~X21, ~ X8, ~X11, ~X12, ~X13, ~X14, ~X22, ~X23, ~ X24.

with means MS well above the global mean of MO = 1.67:

$$Z1 \wedge RX1 \wedge Y1 \qquad \text{(V)}$$

with MS = 2.33, NS = 47, s = 0.793

Accompanying variables: $X15$, ~ $X11$, ~ $X12$, ~ $X13$, ~ $X14$

$$X15 \wedge RY16 \wedge (RX7 \vee RX9) \qquad \text{(VI)}$$

MS = 2.46, NS = 37, s = 0.675

Accompanying variables: $Z1$, ~ $Y1$, ~ $X10$, ~ $X11$, ~ $X12$, ~ $X13$, ~ $X14$, ~ $X24$

Both conjunctions represent large parts of the active compounds and therefore tell the same story as the LOGANA conjunctions characterizing basic features for activity. The development of an already more specific conjunction (VII) is shown in Table 6. Translating back this conjunction to structural terms leads to Figure 5. As was to be expected this structure contains all features already known to be important for very high activity (see Fig. 3) but yields the additional information that the phenyl ring adjacdent to the nitrogen is not *para*-substituted and that the structure $-N = C - (NH_2) - S -$ (variable $X20$) also is typical of high activity compounds. As in conjunction (IV), several X-ring structures are indicated as not being representative for high activity. Conjunction (VII) cannot be interpreted to mean that *para*-substituents, for example, are always unfavorable for high activity (see compound 12) or that phenyl is an "unfavorable" moiety to be placed in X (see compounds 4 and 14). What it in face means is that, within the training series, a set of compounds exists that have the features presented by this conjunction in common and that are all highly active so that it can be regarded a

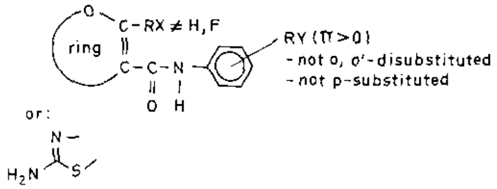

FIGURE 5. Structural pattern resulting from conjunction VII.

reasonable approach to start from these features when synthesizing new compounds.

If the variable $RY1$ is added to conjunction (VII) one obtains:

$$(X16 \vee X20) \wedge RY16 \wedge \sim RY18 \wedge \sim RY19 \wedge RY1 \qquad \text{(VIII)}$$

$MS = 3.42$, $NS = 5$, $s = 0.343$

This conjunction yields an object group with a very high activity mean but is less representative because of the small number of compounds (only five) involved. It may, nevertheless, be taken as an indication that methyl substitution in the *meta* position of *Y* may be another typical feature of the very active compounds. This can be supported by the following simple conjunction which also contains $RY1$ as variable and shows MS well above the overall mean:

$$Y1 \wedge \sim RX12 \wedge RY1 \qquad \text{(XII)}$$

with $MS = 2.64$, $NS = 11$, $s = 0.767$

If one tests other RY substituents in a similar way it can be seen, for example, that the 2-CH_3 group is much less typical of high activity. Thus, in synthesizing compounds supposed to show very high activity it would be a reasonable concept to place a methyl (or a similar) group in the *meta* position of Y but not in the *ortho* or *para* position.

Summarizing Conclusions. Both, LOGANA and LOCON, yield consistent results which allow one to formulate some general rules for high activity as follows:

- Presence of the intact amide group
- Ring in position X (see Figure 1) that has a C=C double bond adjacent to the C atom of the amide moiety
- "Favorable" rings are those which have an oxygen in the upper *meta* position of X (as, e.g., in ring A) with respect to the amide group and, possibly, sulfur in the *ortho* position. A feature to be avoided in X is ring H.
- RX in *ortho* position of X must not be H, F, or phenyl. A typical substituent in this position is CH3. This may be interpreted to mean that substituents of medium size are favorable for high activity
- Phenyl (or, possibly, other aromatic units) in Y
- RY should be hydrophobic and be placed in the *meta* position

These conclusions well agree with the results obtained from an earlier topological analysis based on information theory (*29*) and with results from nonelementary discriminant analysis (*47*) as well as with empirical rules known for the type of antifungal compounds considered. Since in discriminant analysis physicochemical parameters were used (together with indicator variables), a comparison of the conclusions from discriminant analysis with the present results may be of interest. The four most important variables appearing in the discriminant function separating the active compounds into three classes are: f (X), the hydrophobicity of X; σ_m^2(X), where σ_m(X) is the electronic substituent constant characterizing the upper *meta* position (with respect to the amide moiety) in ring X; MR$(RX)_o$, molar refractivity of RX substituents in the *ortho* position; and π_m(RY), the hydrophobicity of RY substituents in the *meta* position. The directionality of these variables is such that increasing values increase the probability that a compound will belong to the most active class.

The first variable indicates that the hydrophobicity of X plays some role, while the second shows that the type of atom in the *meta* position of X is important. The conclusion is the same as from LOGANA and LOCON, namely, that an oxygen in this position should be a good choice. The MR$(RX)_o$ variable indicates the importance of the substituent in *ortho* position of X and of its size, which also clearly came out as one of the most important features from LOGANA and LOCON. In accordance with rule 6 (see above), the π_m(RY) variable, finally indicates that hydrophobic substituents in the *meta* position of the phenyl ring in Y are obviously activity-enhancing. If the separation of active from inactive compounds is investigated by discriminant analysis, the result is identical with that represented by conjunction (1).

It is quite satisfactory that two completely different types of methods using different types of descriptors lead to consistent results. The advantage of LOGANA and LOCON when compared with discriminant analysis in the present case is that the results are obtained in a very straightforward manner directly leading to chemical structures while the complete discriminant function is so complex that it does not admit of a chemical interpretation. Though the patterns found by LOGANA and LOCON provide a good systematization of the compounds in structural terms allowing one to understand why active compounds are active, they would probably not lead to novel structures of outstanding activity. This, however, is not a problem with the method but rather connected with the fact that the training series, though structurally diverse, is still too limited for elucidating structural information which could be regarded as a true surprise. According to the above rules a compound with X = ring A, $RX_o = CH_3$, Z = CONH, Y = phenyl and RY = m-CH_3 should be a very good candidate for high activity. This compound, however, has already been made and is the most active in the training series (No. 1 in Table 1).

Concluding Remarks

The results of this analysis and further examples (*31,32*, and unpublished work) show that LOGANA and LOCON are capable of handling quite diverse structures and provide a clear and consistent picture of structural patterns typical of activity in general and of high activity in particular. Although the data considered are not very extensive, the structural variation in the training series is already too extensive to allow the application of simple linear QSAR models as, for example, Hansch or Free-Wilson analysis. None of these methods gives satisfactory results in the present case so that a topological analysis is the only choice. LOGANA and LOCON have the following advantages:

- No assumptions about additivity of effects or linear models are necessary.
- The contribution of a given structural feature to the biological activity need not be a consistent factor.
- A uniform mechanism of action for all compounds of the training series is not required. If different mechanisms are operating they will be reflected by different pharmacophores provided, of course, that a representative number of compounds exists for each mechanism.
- There is no dependence on data distribution or statistical formalisms.
- Full use is made of the intuition and experience of the researcher.
- The results are directly obtained in terms of a structural pattern.
- Physicochemical molecular parameters are usually not needed. Only when the variation of substituents becomes an important property are substituent constants required to parametrize substituents via binary descriptors (*31,32*). Even then, however, approximate values are sufficient.
- The methods can be applied to extremely diverse data sets where linear free energy formalisms are bound to fail. Inactive compounds can be included in the calculation.
- For LOGANA, biological activity need not be available on a continuous scale. A simple classification is sufficient so that it is possible to use data from mass screening or sampled over a long period of time or even from different sources.

Most specific for LOGANA and LOCON are the second through sixth points above, while the others are shared by the majority of the other typological methods mentioned in a previous section. It should be mentioned in this context that the STRAC procedure also uses the interactive construction of conjunctions by logical operations with decision-making by the operator. The objective, however, is different from LOGANA in that not a topological pattern but the evaluation of discriminating features with the final purpose of classifying compounds via probability-derived criteria is aimed at. Other differences are that STRAC cannot handle very large and diverse data sets and that it is directly at lead optimization while LOGANA and LOCON are more of the lead generation type.

Which of the topological procedures available and what kind of features are to be applied in a particular case depend on the data and the objective of the analysis. LOGANA and LOCON are the methods of choice when not only the type of substructures but also the molecular region where they occur can be (at least loosely) coded for and when the primary purpose of the analysis is to evaluate basic principles of a given type of biological activity in terms of chemical topology. Once these principles are known the action of compounds may be better understood and systematized, and it becomes possible to design new structures possessing the desired type of biological activity. This also includes the possibility to design compounds which are not likely to possess an undesired side effect (e.g., a certain type of chemical hazard) when the corresponding topological pattern has been evaluated. The design space can be extended by introducing the concept of bioisosterism in a similar manner as, for example, in Magee's RANGE procedure (*48*).

When simple classification or preselection of already existing compounds and not so much interpretation and design are in the foreground the "index" or classification methods outlined above are to be applied. Such methods have also been applied in the field of chemical hazards such as toxic effects (*49,50*) carcinogenicity (*20,49–52*), and mutagenicity (*43,44,49*) not without success, although the results thus far obtained more or less suffer from the deficiencies discussed earlier (e.g., extremely complex classifier). Considering the number of chemicals already in circulation and those to be expected from further syntheses, such analyses are to be regarded as indispensable tools to set priorities for biological testing. It is important to be well aware of their limitations which, in particular, means that predicted data should never be taken for granted and must not be allowed to replace experimental measurements (*53*).

When applying topological methods one must, of course, keep in mind several restrictions and possible pitfalls partly already discussed. As in any QSAR method the properties of the training series are crucial for the validity of the results. Structural variations not adequately represented in the training series can also not adequately be reflected by the results. For LOGANA and LOCON, for example, this has the consequence that the patterns evaluated cannot be regarded as a general or final solution. All that can be said is that they are typical of the highly active analogs in that part of the chemical compound space that is covered by the training series and that there is a high probability of finding other highly active compounds if these patterns are used as a design criterion.

A very serious limitation of all topological approaches results from the fact that the actual events of drug–receptor interactions are three-dimensional and dynamic, while these methods are static and operate in only two dimensions. Topological approaches will, therefore, yield valid results only to the extent that two-dimensional chemical structures reflect the much more complicated processes operating when drugs interact with biological targets. This implies that in a number of cases such methods are bound to fail because the topological descriptions of structures is inadequate; even if geometrical descriptors are added (*8*), which can, in principle, be done in all methods, this would still be true because of the static kind of approach. Molecular modeling techniques are then the most promising way to go but even in connection with such methods topological analyses can be very valuable since their results may provide hypotheses which can be used as an input when applying the more sophisticated techniques.

REFERENCES

1. Purcell, W. P., Bass, G. E., and Clayton, J. M. Strategy in Drug Design, A Molecular Guide to Biological Activity. Wiley-Interscience, New York, 1973.
2. Tichý, M., Ed. Quantitative Structure-Activity Analysis. Akademiai Kiado, Budapest, 1976.
3. Buisman, J. A. K., Ed. Biological Activity and Chemical Structure, Pharmacochemistry Library, Vol. 2. Elsevier, Amsterdam, 1977.
4. Martin, Y. C. Drug Design Methods: A Critical Introduction. Marcel Dekker, New York, 1978.
5. Franke, R., and Oehme, P., Eds. Quantitative Structure-Activity Analysis, Akademie-Verlag, Berlin, 1978.
6. Golender, V. E., and Rozenblit, A. B. Computer Assisted Methods of Drug Design. Zinatne, Riga, 1978 (in Russian).
7. Seydel, J. K., and Schaper, K. J. Chemische Struktur und biologische Aktivitaet von Wirkstoffen. Verlag Chemie, Weinheim, 1979.
8. Stuper, A. J., Bruegger, W. E., and Jurs, P. C. Computer Assisted Studies of Chemical Structure and Biological Function. Wiley, New York, 1979.
9. Olson, E. C., and Christoffersen, R. E., Eds. Computer-Assisted Drug Design. ACS Symposium Series 112, American Chemical Society, Washington, DC, 1979.
10. Knoll, J., and Darvas, F., Eds. Chemical Structure–Biological Activity Relationships. Akademiai Kiado, Budapest, 1980.
11. Franke, R. Optimierungsmethoden in der Wirkstofforschung. Akademie-Verlag, Berlin, 1980.
12. Dearden, J. C., Ed. Quantitative Approaches to Drug Design. Pharmacochemistry Library, Vol. 6. Elsevier, Amsterdam, 1983.
13. Topliss, J. G., Ed. Quantitative Structure-Activity Relationships of Drugs. Academic Press, New York, 1983.
14. Kuchar, M., Ed. Quantitative Structure-Activity Relationships in Design of Bioactive Compounds. J. A. Prous, S. A., Barcelona, in press.
15. Franke, R. Theoretical Drug Design Methods. Pharmacochemistry Library, Vol. 7. Elsevier, Amsterdam, 1984.
16. Leffler, J. E., and Grunewald, E. Rates and Equilibria of Organic Reactions. Wiley, New York, 1963.
17. Adamson, G. W., and Bawden, D. A. Substructural analysis methods for structure-activity correlation of heterocyclic compounds using Wiswesser line notation. J. Chem. Inf. Comput. Sci. 17: 164–171 (1977).
18. Bawden, D. Computerized chemical structure-handling techniques in structure-activity studies and molecular property prediction. J. Chem. Inf. Comput. Sci. 23: 14–22 (1983).
19. Smith, E. G., and Baker, P. A. The Wiswesser Line-Formula Chemical Notation, 3rd ed. Chemical Information Management Inc., Cherry Hill, NJ, 1975.
20. Jurs, P. C., Chou, J. T., and Yuan, M. Computer-assisted structure-activity studies of chemical carcinogens. A heterogeneous data set. J. Med. Chem. 22: 476–483 (1979).
21. Kirschner, G. L., and Kowalski, B. R. The application of pattern recognition to drug design. In: Drug Design, Vol. VIII (E. J. Ariens, Ed.), Academic Press, New York, 1979, pp. 73–131.
22. Golender, V. E., and Rozenblit, A. B. Logico-structural approach to computer-assisted drug design. In: Drug Design, Vol. IX (E. J. Ariens, Ed.), Academic Press, New York, 1980, pp. 299–337.
23. Avidon, V. V., Pomerantsev, I. A., Golender, V. E., and Rozenblit, A. B. Structure-activity oriented languages for chemical structure representation. J. Chem. Inf. Comput. Sci. 22: 207–214 (1982).
24. Rozenblit, A. B. Computer–assisted drug design. Strategy and algorithms. In: Strategy in Drug Design (J. A. Keverling Buisman, Ed.), Elsevier, Amsterdam, 1982, pp. 287–307.
25. Cammarata, A., and Menon, G. K. Pattern recognition classification of therapeutic agents according to pharmacophore. J. Med. Chem. 19: 739–748 (1976).
26. Menon, G. K., and Cammarata, A. Pattern recognition. II. Investigation of structure-activity relationships. J. Pharm Sci. 66: 304–314 (1977).
27. Henry, D. R., and Block, J. H. Classification of drugs by discriminant analysis using fragment molecular connectivity. J. Med. Chem. 22: 465–472 (1979).
28. Henry, D. R., and Block, J. H. Pattern recognition of steroids using fragment molecular connectivity. Eur. J. Med. Chem. 15: 133–138 (1980).
29. Huebel, S., Roesner, T., and Franke, R. The evaluation of topological pharmacophores by heuristic approach. Pharmazie 35: 424–433 (1980).
30. Franke, R., Streich, W. J., and Huebel, S. Topological pharmacophores from continuous biological activity data. Studia Biophys. 97: 11–20 (1983).
31. Franke, R., and Streich, W. J. Topological pharmacophores: new methods and their application to a set of antimalarials. 2. Results from LOGANA. Quant. Struct. Act. Relat., in press.
32. Franke, R., and Streich, W. J. Topological pharmacophores: new methods and their application to a set of antimalarials. 3. Results from LOCON. Quant. Struct. Act. Relat., in press.
33. Mercer, C., and Dubois, J. E. Comparison of molecular connectivity and DARC/PELCO methods: performance in antimicrobial halogenated phenol QSARs. Eur. J. Med. Chem. 14: 415–423 (1979).
34. Balaban, A. T., Chiriac, A., Motoc, I., and Simon, Z. Steric Fit in Quantitative Structure-Activity Relations. Springer-Verlag, Berlin, 1980.
35. Kier, L. B., and Hall, L. H. Molecular Connectivity in Chemistry and Drug Research. Academic Press, New York, 1976.
36. Adamson, G. W., and Bawden, D. An empirical method of structure-activity correlation for polysubstituted cyclic compounds using Wiswesser line notation. J. Chem. Inf. Comput. Sci. 16: 161–165 (1976).
37. Adamson, G. W., and Bawden, D. Substructural analysis techniques for empirical structure-property correlation. Application to stereochemically related molecular properties. J. Chem. Inf. Comput. Sci. 20: 97–100 (1980).
38. Cramer, R. D., and Redl, G., and Berkoff, C. E. Substructural analysis. A novel approach to the problem of drug design. J. Med. Chem. 17: 533–535 (1974).
39. Hodes, L., Hazard, G. F., Geran, R. I., Richman, S. A statistical-heuristic method for automatic selection of drugs for screening. J. Med. Chem. 20: 469–475 (1977).
40. Hodes, L. Computer-aided selection of novel antitumor drugs for animal screening. In: Computer-Assisted Drug Design (E. C. Olsen and R. E. Christopherson, Eds.), ACS Symp. Ser. 112, American Chemical Society, Washington, DC, 1979, pp. 583–602.
41. Hodes, L. Computer-aided selection of compounds for antitumor screening: validation of a statistical-heuristic method. J. Chem. Inf. Comput. Sci. 21: 128–132 (1981).
42. Hodes, L. Selection of molecular fragment features for structure-activity studies in antitumor screening. J. Chem. Inf. Comput. Sci. 21: 132–136 (1981).
43. Tinker, J. F. Relating Mutagenicity to Chemical Structure. J. Chem. Inf. Comput. Sci. 21: 3–7 (1981).
44. Tinker, J. F. A computerized structure-activity correlation program for relating bacterial mutagenesis activity to chemical structure. J. Comput. Chem. 2: 231–243 (1981).
45. Streich, W. J., and Franke, R. Topological pharmacophores: new methods and their application. 1. The methods LOGANA and LOCON. Quant. Struct. Act. Relat. 4: 13–18 (1985).
46. White, G. A., and Thorn, G. D. Structure-activity relationship for carboxamide fungicides and succinic-dehydrogenase complex

of *Cryptococcus laurentii* and *Ustilago maydis*. Pestic. Biochem. Physiol. 5: 380–395 (1975).

47. Dove, S., and Franke, R. Discriminant analysis and QSAR work. In: Quantitative Structure-Activity Analysis (R. Franke and P. Oehme, Eds.), Akademie-Verlag, Berlin, 1978.
48. Magee, P. S. A new approach to bioactive synthesis. In: Computer-Assisted Drug Design (E. C. Olsen and R. E. Christopherson, Eds.), ACS Symp. Ser. 112, American Chemical Society, Washington, DC, 1979, pp. 319–340.
49. Enslein, K., and Craig, P. N. A toxicity estimation model. J. Environ. Pathol. Toxicol. 2: 115–121 (1978).
50. Enslein, K., and Craig, P. N. Status Report on Development of Predictive Models of Toxicological Endpoints. Genese Corp., Rochester, 1979.
51. Yuta, K., and Jurs, P. C. Computer-assisted structure-activity studies of chemical carcinogens. Aromatic amines. J. Med. Chem. 24: 241–251 (1981).
52. Chou, J. T., and Jurs, P. C. Computer-assisted structure-activity studies of chemical carcinogens. An *N*-nitroso compound data set. J. Med. Chem. 22: 792–797 (1979).
53. Rekker, R. F. LD50 values: are they about to become predictable? TIPS 383–384 (1980).